Naval Research Laboratory

Washington, DC 20375-5320

NRL/FR/5550--01-10,016

Variable Data Rate Voice Encoder for Voice Over Internet Protocol (VoIP)

GEORGE S. KANG

Transmission Technology Branch
Information Technology Division

December 28, 2001

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
December 28, 2001	Formal	October 1, 2000 to August 30, 2001

4. TITLE AND SUBTITLE

Variable Data Rate Voice Encoder for Voice Over Internet Protocol (VoIP)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
33904N, 61553N

6. AUTHOR(S)

George S. Kang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/FR/5550--01-10,016

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commander
Space and Naval Warfare Systems Command
4301 Pacific Highway
San Diego, California 92110-3127

10. SPONSOR / MONITOR'S ACRONYM(S)

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Over the years, all DOD voice terminals transmitted speech at a constant data rate. That was the era when each user was allotted a fixed amount of channel resources (e.g., bandwidth). Now, the age of Voice over Internet Protocol (VoIP) has arrived. In this transmission approach, the channel resource is shared by all users, and speech data is transmitted in packets in which packet size can be varied. Hence, speech can be transmitted at a variable data rate (VDR).

The author recognized that speech is a VDR information source where the speech waveform is a time-varying mixture of complex waveforms (vowels) and simple waveforms (consonants and non-speech gaps). The author combined the VDR nature of speech and the VDR capability of VoIP to achieve more efficient voice transmission while achieving higher speech quality.

The author's voice encoding technique supports the future Network Centric Warfare described in Joint Vision 2010. Recognizing the significance of this R&D product, the Navy Information Assurance Program Management Office (SPAWAR PMW161) will Fleet-test the author's VDR voice encoder during FY03.

15. SUBJECT TERMS

Variable data rate vocoder, speech modeling, residual-excited LPC

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					George S. Kang
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UL	24	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified			202-767-2157

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

CONTENTS

VARIABLE-DATA-RATE VOICE ENCODER FOR VOICE OVER INTERNET PROTOCOL (VOIP)

INTRODUCTION

Currently, the Department of Defense (DOD) and private industries have been developing technologies to transmit speech over the Internet. The Naval Research Laboratory (NRL) is also developing an Internet Protocol (IP) voice terminal with a gateway that permits interoperation between the Voice over Internet Protocol (VoIP) and DOD legacy voice equipment. The variable-data-rate (VDR) voice processor documented in this report is for this NRL effort to transmit voice efficiently (i.e., higher speech quality at a lower data rate) for VoIP so that the network can be used by more users.

Historically, voice communication has taken place at a constant data rate, and each communicator is allotted a channel with fixed bandwidth. If the user has a wideband channel (having a 25-kHz bandwidth), speech can be transmitted at as high a data rate as 32 kilobits per second (kb/s), and speech quality is excellent. Voice transmission, however, is inefficient at that high a data rate. Because the speech waveform is a time-varying mixture of complex waveforms (vowels) and simple waveforms (consonants and interword or interphrase gaps), that high a data rate is not needed at all times to encode speech.. On the other hand, if the user has a narrowband channel (having a 3-kHz bandwidth), voice data rate is typically 2.4 kb/s. At this low data rate, speech quality has not been acceptable to all users. With some amount of acoustic noise in the background or with a high-pitched female voice as the input, encoded speech will not be sufficiently intelligible for DOD applications.

This is the age of VoIP. Significant VoIP characteristics include the following:

- all users share the network resources (i.e., no one gets a fixed allotment of channel resources),
- speech is transmitted in small groups of data called packets, and
- the packet size can be variable. In other words, speech can be transmitted at as high as 32 kb/s to produce high-quality vowels or as low as a few hundred bits per second to produce good-sounding consonants or background noise during silence.

Given a large aggregate of users typically sharing network resources, one user s larger data rate is offset by another user s smaller data rate. Therefore, the sum of all users throughputs remains more or less at a constant level.

To make the VoIP transmission efficient and to allow more network users, we developed a VDR voice processor. There has been no reported VDR voice processor elsewhere having the merits (presented below) of our VDR voice processors. We have posted three audio clips on our website and have included them on the CD attached to this report to demonstrate the audio qualities of our VDR processor described below.

- Reduction of Average Data Rate in the Event of Network Congestion: Our VDR voice processor provides four different average data rates: 9.6, 12, 16, and 20 kb/s. On the basis of the network traffic condition, we may select a preferred average data rate. For any average

Manuscript approved October 17, 2001.

data rate, seven instantaneous data rates are generated 44 times per second (the frame rate identical to that of the DOD narrowband tactical voice terminals). Based on the complexity of the speech waveform, an appropriate data rate is selected automatically. Essentially, our VDR voice processor is four VDR voice processors combined into one.

- High speech quality: The speech quality at an average data rate of 20 kb/s compares favorably to the speech encoded at a constant data rate of 64 kb/s. As average data rates decrease from 20 kb/s to 9.6 kb/s, speech quality degrades only slightly.

- Data rate switching does not generate clicks or pops in synthesized speech: We use a single speech processing principle to generate a single set of speech parameters. This single set of speech parameters is quantized (i.e., encoded) differently to produce variable data rates, but all these data rates are synchronized together. Therefore, the synthesized speech does not produce clicks or pops due to waveform discontinuity at the frame boundary. Even if the average data rate is switched during continuous speech, there will be no audible clicks. This is the most remarkable aspect of our VDR voice processor.

- Acoustic noise tolerance: Our VDR voice processor produces speech that is highly tolerant to acoustic noise interference, as demonstrated by the audio samples. This property is highly desirable for military voice communication.

- No cropping of silence waveform: During periods of silence, we don t need to crop the speech waveform to make the average data rate lower because the data rate is automatically reduced. Cropping of the speech waveform can be difficult for softly spoken speech or noisy speech, and it is also difficult to fill the gaps at the receiver with the waveform that is acoustically compatible with background noise. Even with these technical difficulties aside, cropping of the silent waveform is a bad idea for tactical voice communication because background noise (from such sources as gunshots, explosions, aircraft, and voices heard in the background) provides vital tactical information in conducting warfare. With our VDR voice processor, the data rate is automatically reduced to a minimum during silent periods but is large enough to produce realistic background sounds at the receiver to gauge activity at the transmitter site.

- Direct interoperability with DOD narrowband vocoders: We designed our VDR voice processor to interoperate directly with DOD tactical legacy voice equipment operating at a low data rate of 2.4 kb/s. Therefore, we intentionally exploited the voice data used by the legacy equipment in generating variable data rates. Then we intentionally embedded this voice data into the voice data of our VDR voice processor. As a result, our VDR voice processor interoperates directly with Advanced Narrowband Digital Voice Terminals (ANDVT) and Secure Telephone Units (STU-III, 2.4 kb/s mode).

Conserving by minimizing the voice data rate is always a good principle to follow in the DOD voice communication design, but not when it sacrifices speech intelligibility. Speech is a real-time phenomenon that does not give us a chance for rehearing. We must understand the meaning of the received speech at the first hearing. A misunderstanding or a failure to understand tactical messages could bring a disastrous consequence. Our VDR voice processor generates high-quality speech at any of the four average data rates mentioned above. This yields a reduced chance of misunderstanding the received speech at the first hearing.

This report is the result of our continued efforts to improve naval secure voice communication. It is fortunate that this R&D effort is sponsored in part by the Navy Information Systems Security Program Office (SPAWAR, PMW-161), which procures and deploys new and improved secure voice capabilities for the Fleet. A Fleet-test of the VDR voice processor documented in this report is already planned.

BACKGROUND

Circuit Switching

Since the invention of the telephone in 1876 until recent years, a two-way telephone conversation was exchanged over the switched circuits allotted exclusively for the users (Fig. 1). From the 1950s, DOD voice communication has been increasingly encoded by digital, rather than analog, techniques because encryption of voice data is both easier and more secure. Yet, the basic approach to channel resources has not changed from the days of analog speech transmission; circuit switching prevailed through the years. Thus, each user is allotted a channel having a fixed bandwidth. Since channel bandwidth is fixed, speech is transmitted at a constant and maximum data rate compatible with the allotted channel bandwidth.

All current DOD secure phones still operate at a fixed rate. Examples are the ANDVT operating at 2.4 kilobits per second (kb/s), the STU-III operating at 2.4 or 4.8 kb/s, the Single Channel Ground and Airborne Radio System (SINCGARS) operating at 16 kb/s, and more recently, the Secure Telephone Equipment (STE) operating at 32 kb/s. Currently, tens of thousands of these secure telephones are supporting tactical and office-to-office voice communications effectively. They will not fade away any time soon. Any new DOD voice encoder, such as our VDR voice processor, must be designed to interoperate with this legacy voice equipment.

Fig. 1 — Example of circuit switching. Telephone switchboards came into use at the turn of the 20th century. This marked the beginning of circuit switching, which has remained in use for the last 100 or more years. In this approach, each user is given a dedicated circuit for the duration of communication. Late in the second decade of the 20th century, automated circuit switching began to phase in.

Packet Switching

In the 1980s, the DOD Advanced Research Projects Agency (DARPA) promoted R&D efforts to develop packetized communication techniques to achieve more efficient utilization of network resources. Packet switching has unique features not found in circuit switching, such as:

- Network resources are shared by all users all the time (Fig. 2);
- Data transmission is in packets (small groups of data);
- Each packet may be variable in length and duration [1]; and
- Each packet is delivered via various routes if the networks are interconnected.

If the packet size is variable, then the data rate is variable, provided that the frame rate is constant. Fortunately, speech is a variable data rate source, as discussed in the next section. We prefer a constant frame rate because all different data rates are synchronized and constant switching of data rates does not introduce clicks or pops in the synthesized speech.

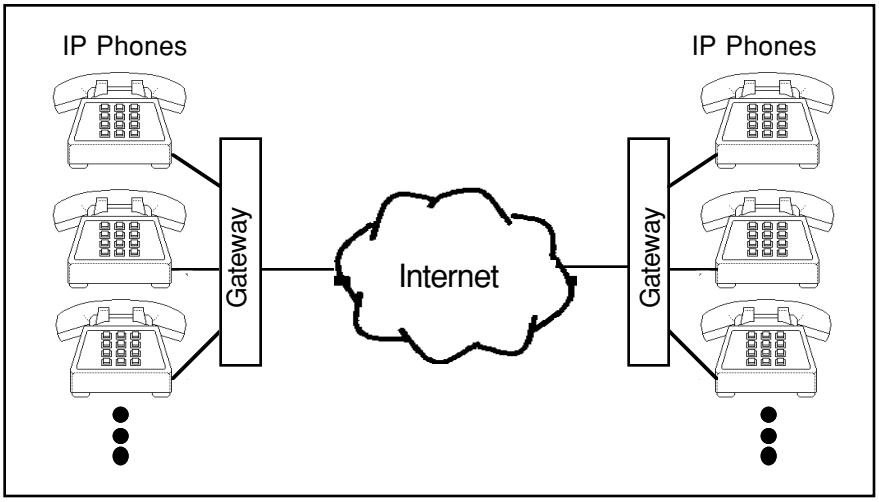


Fig. 2 — Example of packet switching. A significant feature of packetized communication is that the channel resources are shared by all users. Thus, one user s high data rate is offset by another user s low data rate. Hence, everyone can transmit higher quality speech while keeping the total throughput of the network more or less unchanged.

Resource sharing has been practiced in other fields. One example is the electric power distribution system delivering power to each household. As we know, while one consumer may be using a large amount of electric power, another consumer may be using little power during the same time period. Hence, consumers can meet their individual power needs while maintaining the aggregate of the total power within the allowable capacity of the distribution system. Resource sharing benefits all users.

Speech is a Variable Data Rate Source

Fortunately, speech is a time-varying mixture of complex and simple waveforms. It is a result of continuous dynamic activity of the mouth for generating vowels, interrupted at explosive sounds (/t/, /p/, /k/, etc.) or breath control demand. It generates a stream of sounds having variable spectral contents. Hence, the data rate required to represent speech sounds varies rapidly. This makes packetized switching ideally suited for voice transmission.

As exemplified in Fig. 3, a vowel has a complex waveform because it has three to four resonant frequencies. In addition, the vowel waveform is modulated by a pitch frequency (i.e., a vowel waveform is repetitive) that must be reproduced accurately. Therefore, a higher data rate of 30 kb/s or more is needed to represent a vowel accurately. If the data rate is lower, the synthesized speech of a vowel will sound raspy, fluttery, and wobbly. Human ears are particularly sensitive to those speech impediments because they are absent in natural speech.

On the other hand, the consonant waveform (for /s/, /sh/, /t/, /p/, etc.) has typically one weak resonant frequency and it is not pitch-modulated (i.e., not repetitive). We note that the sound quality of a consonant varies significantly from one speaker to another. Hence, a consonant waveform can be encoded crudely without being misunderstood. A lower data rate of a few kb/s is adequate to describe a consonant.

Speech also has gaps within words, between words, and between phrases. Our VDR voice encoder still transmits whatever the waveform presents during the gap because those waveforms may include gunshot, explosion, airplane, or other background noise. The noise provides important tactical information to warfighters. We need only a few hundred bits to describe those nonspeech sounds.

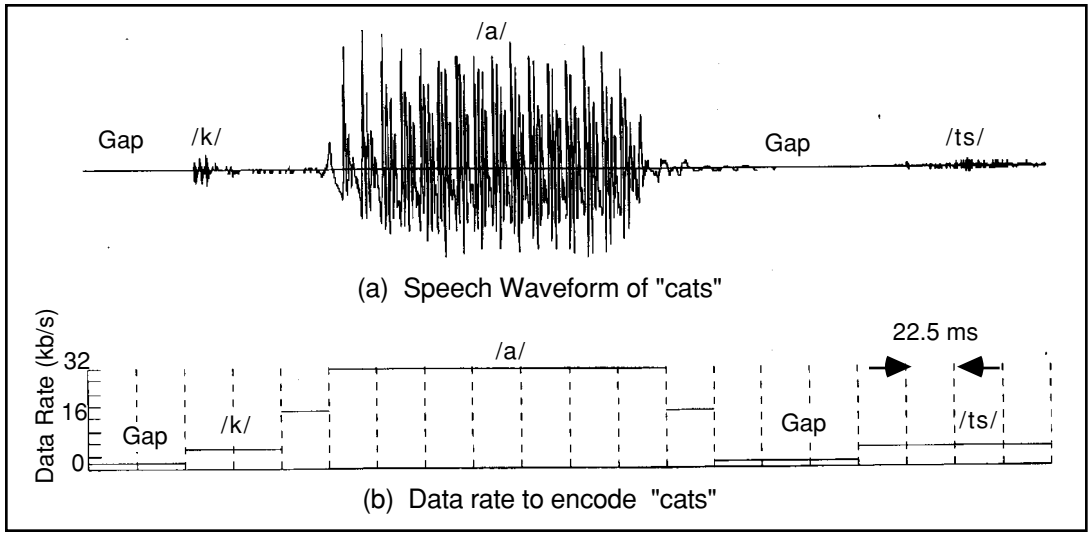


Fig. 3 — Speech waveform of (a) cats and (b) the data rate required to encode cats. As noted in Fig. 3(b), vowel waveform /a/ requires a data rate of as much as 32 kb/s to generate high quality vowel, whereas consonants /k/ and /ts/ require a data rate of approximately 8 kb/s to generate the speech quality. To encode gaps requires less than 500 bits/second. This example shows that encoding of speech at a constant rate of 32 kb/s is a waste of data. The average data of this example is 20 kb/s. The time mark of 22.5 ms is the frame size during which the speech parameters are generated and encoded.

Use of the Single Voice Processing Principle

Our VDR voice encoder is not a collection of various vocoders operating at different data rates. If it were, we could not exploit the variable-data-rate nature of the speech waveform mentioned in the preceding section. Furthermore, it would be impossible to switch data rates during speech without introducing pops and clicks, or even undesirable cropping of speech.

In our VDR voice processor, we use a single voice processing principle to generate four different average data rates: 9.6, 12, 16, and 20 kb/s, one of which is selected based on the network traffic condition. We preselected seven instantaneous data rates (i.e., data rates from each frame of 22.5 ms) to represent the simplest speech waveform to most complex speech waveform. We could use more than seven instantaneous data rates but the resultant benefit is negligible because we cannot perceive a change in speech quality if the data rate increment is too small. Figure 4 shows the four sets of seven instantaneous data rates. As noted, each set of instantaneous frequencies is clustered differently to provide a different average data rate.

These instantaneous data rates are generated from one common set of speech parameters generated in each frame. These speech parameters are quantized differently to produce 28 different instantaneous data rates. Since all these data rates are synchronized, we can switch average data rate even during continuous speech, without generating clicks or pops caused by waveform discontinuities present at the boundary of frames.

Our highest average data rate is 20 kb/s. According to our formalized intelligibility testing and also extensive listening tests, speech quality at an average data rate of 20 kb/s compares favorably with that of a fixed data rate of 64 kb/s. In other words, we don t need an average data rate higher than 20 kb/s. In addition, we limit the lowest average data rate to 9.6 kb/s. These prescribed limits of VDR data rate produce speech quality judged to be acceptable to all users even under difficult operating conditions (i.e., severe acoustic noise interference or extremely high-pitched female voice, etc.).

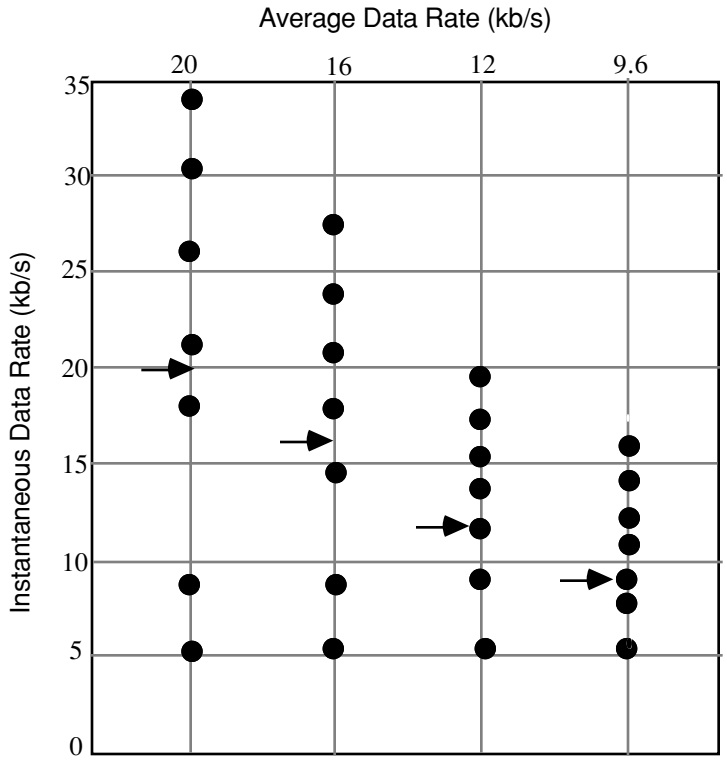


Fig. 4 — Instantaneous data rates to produce four different average data rates are indicated by arrows. For each frame of 22.5 ms, a set of speech parameters is generated. They are quantized 28 different ways, resulting in seven instantaneous data rates in four different average data rate groups.

Prior Variable Data Rate Effort

In 1996, a variable data voice processor was implemented [2] based on the Linear Predictive Coder (LPC) operating at 2.4 kb/s, as used in narrowband tactical voice terminals such as ANDVT and STU-III [3]. There are three different speech conditions in which speech data per frame are varied:

- When speech is voiced (vowels), the output bit stream is 54 bits/frame, identical to that of the 2.4-kb/s LPC.
- If speech is unvoiced, the output bit stream is reduced to 34 bits/frame by eliminating less significant bits of speech data.
- For periods of silence, no speech data are transmitted.

The average data rate for typical conversational speech is 1 to 1.2 kb/s. This earlier VDR voice processor is developed for low-data-rate applications, whereas the author s VDR voice processor is for medium-data-rate applications. This produces higher quality speech over, for example, shipboard circuits that currently use the STE telephone service.

VDR VOICE PROCESSOR

Efficient speech encoding until now has always relied on a speech analysis/synthesis system in which the speech waveform is decomposed into sets of slow time-varying (or narrowband) parameters and fast time-varying (or wideband) parameters. Narrowband speech parameters are updated once per frame (22.5 ms), whereas wideband speech parameters are updated every speech sampling time interval (0.125 ms). The sum of these data rates is often substantially lower than the speech data rate associated with direct encoding of the speech waveform.

Currently, the speech analysis/synthesis system is used in all voice encoders that operate at a fixed data rate of 32 kb/s or below. Our VDR voice processor also relies on the speech analysis/synthesis system. The basic difference between a fixed rate voice encoding system and a VDR voice encoding system rests on how speech parameters are quantized (i.e., encoded), not how they are extracted.

LPC Analysis/Synthesis System

From among the many different speech analysis/synthesis systems known, we selected LPC analysis/synthesis because we want our VDR voice processor to interoperate directly with DOD narrowband legacy voice equipment that also uses the LPC analysis/synthesis system. The LPC analysis system decomposes the speech waveform into slow time-varying components (10 prediction coefficients, pitch period, and pitch prediction coefficients) and fast time-varying components (pitch-filtered prediction residual). By quantizing each component efficiently, the overall data rate can be lowered by more than what is attainable without relying on the speech analysis/synthesis method. Figure 5 is a block diagram of the LPC analysis/synthesis system.

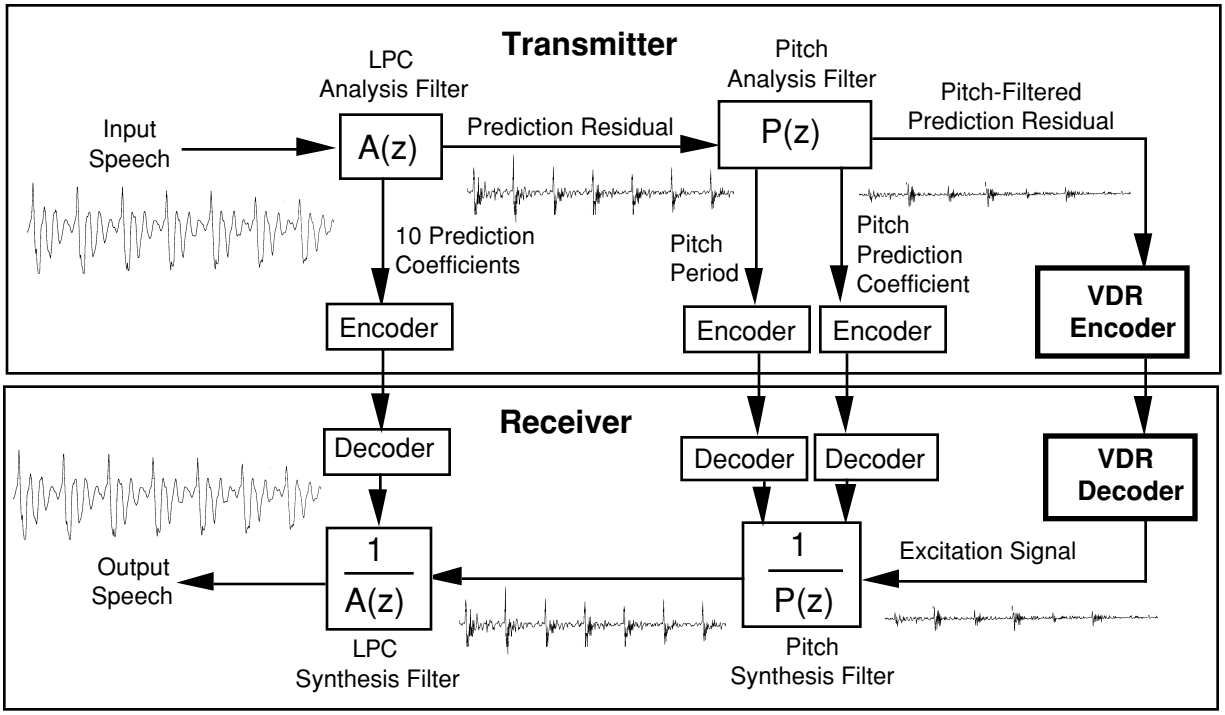


Fig. 5 — Block diagram of LPC-based speech analysis and synthesis. The transmitter is a two-stage spectral flattening process. The receiver is the inverse of the transmitter. The most critical part of the VDR voice processor is the encoder for the pitch-filtered prediction residual, which is the excitation signal to the speech synthesizer at the receiver.

LPC Analysis

The LPC analysis generates two types of prediction analyses: short-term and long-term. A short-term analysis prediction generates 10 LPC coefficients, which provides a means to estimate the speech spectral envelope, whereas a long-term analysis prediction is a means to estimate pitch harmonics that are inscribed under the spectral envelope, as illustrated later.

Short-Term Prediction Analysis

In the short-term prediction analysis, a speech sample is represented by a linear combination of past samples, each separated by a sampling time interval. Thus,

$$x(i) = \sum_{j=1}^{n} \alpha(j) x(i-j) + \varepsilon_s(i) \qquad \text{i = 1, 2, 3, ...,} \tag{1}$$

where $x(i)$ is the ith speech sample, $\alpha(j)$ is the jth prediction coefficient, and $\varepsilon_s(i)$ is the ith short-term prediction residual sample. The quantity n is normally 10. The ith short-term prediction residual sample obtained from Eq. (1), is expressed by

$$\varepsilon_s(i) = x(i) - \sum_{j=1}^{n} \alpha(j) x(i-j). \tag{2}$$

In terms of z-transform, Eq. (2) may be expressed by

$$\begin{aligned} E_S(z) &= \left[1 - \sum_{j=1}^{n} \alpha(j) z^{-j}\right] x(z) \\ &= \left[1 - A_n(z)\right] x(z). \end{aligned} \tag{3}$$

The quantity $\left[1 - A_n(z)\right]$ is the transfer function of the LPC analysis filter that transforms the speech waveform to short-term prediction residual. The frequency response of $\left[1 - A_n(z)\right]$ approximates the inverse of the speech spectral envelope as shown by Fig. 6(b). The short-term prediction residual is considerably free from speech resonance, as demonstrated by a lack of ringing in the waveform in each pitch cycle (see Fig. 5).

Long-Term Prediction Analysis

The short-term prediction residual is predominantly the pitch harmonics that will be removed by the long-term prediction analysis. It is similar to the short-term prediction analysis discussed previously except that the correlation period is the pitch period T (anywhere from 20 to 160 speech sampling time intervals) with only one prediction coefficient. Thus, the long-term prediction residual or pitch-filtered prediction residual is expressed by

$$\varepsilon_l(i) = \varepsilon_s(i) - \beta \varepsilon_s(i-T), \tag{4}$$

where $\varepsilon_l(i)$ and $\varepsilon_s(i)$ are, respectively, the i^{th} long-term and short-term prediction residuals. In terms of z-transform, Eq. (4) may be written as

$$\begin{aligned} E_L(z) &= \left[1-\beta z^{-T}\right] E_S(z) \\ &= [1-P(z)] E_S(z) \end{aligned}, \tag{5}$$

where the quantity $[1-P(z)]$ is the transfer function of the pitch analysis filter, which is a filter to attenuate pitch harmonics (see Fig. 7).

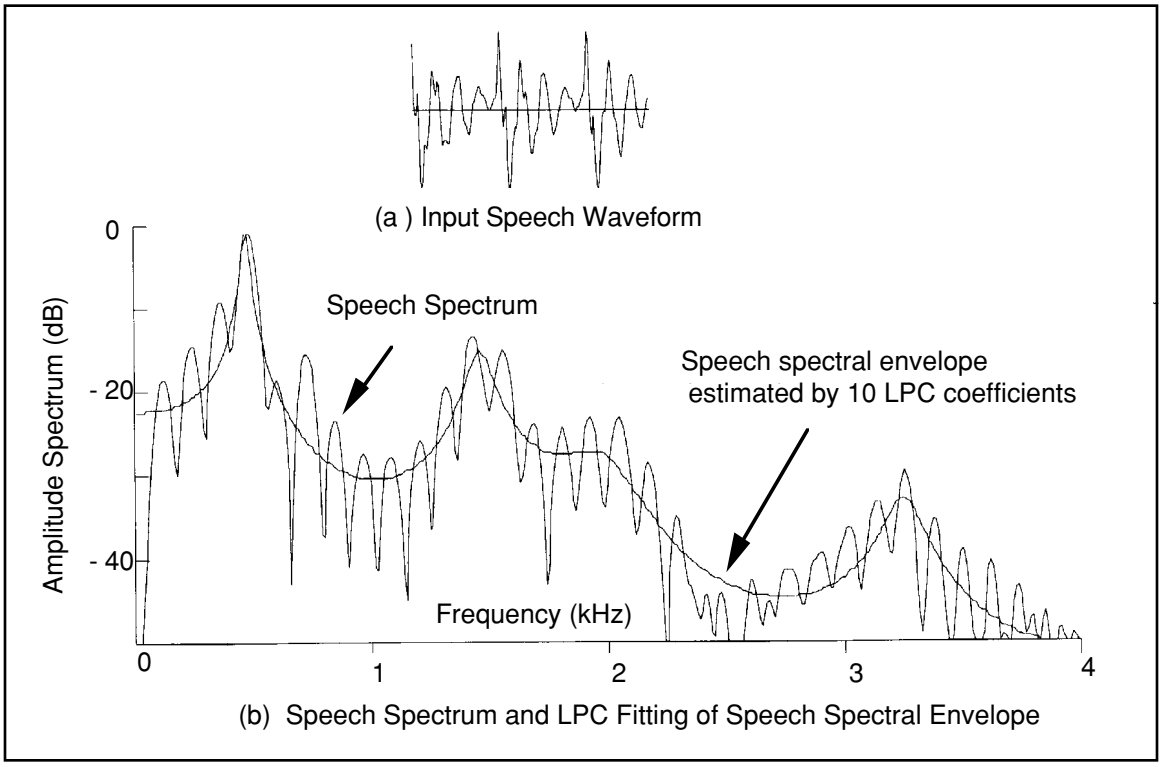


Fig. 6 — Speech spectrum and the speech spectral envelope estimated by the LPC analysis. With 10 LPC coefficients, the estimated speech spectral envelope is good. The difference between the spectrum and the estimated speech spectral envelope is the spectrum of the prediction residual. As shown in Fig. 5, the prediction residual is free from ringing, signifying that the resonant frequencies have been removed.

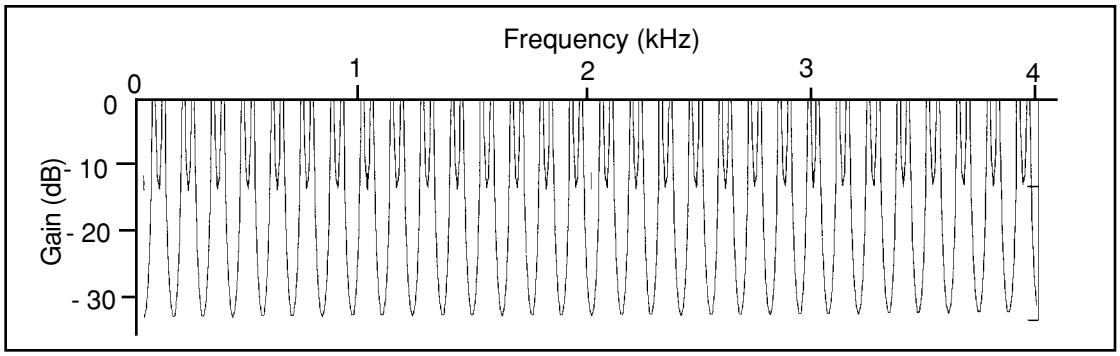


Fig. 7 — Frequency response of the pitch analysis filter. This is a notch filter made up of a single-tap feed-forward filter. This filter suppresses pitch harmonics.

Combining Eqs. (3) and (5), the output of the LPC analysis system in terms of speech is expressed by

$$E_L(z) = [1 - A(z)][1 - P(z)]x(z). \tag{6}$$

The quantity $[1 - A(z)][1 - P(z)]$ is the transfer function of the LPC analysis system having both the short-term and long-term predictions. This LPC analysis filter converts the speech signal into the pitch-filtered prediction residual, which has a flat spectrum as shown by Fig. 8. Quantization of the excitation signal is a major design issue of the VDR voice processor. This issue is discussed in the section on excitation signal quantization.

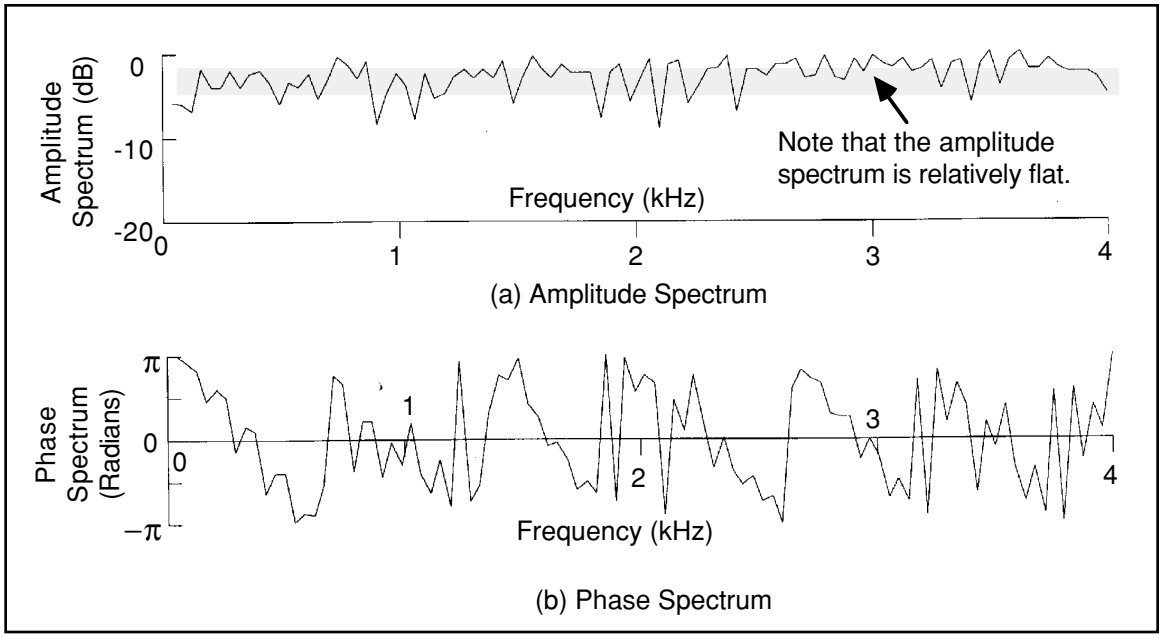


Fig. 8 — Amplitude and phase spectra of the pitch-filter prediction residual (the LPC analysis output). Upon quantization, this signal becomes the excitation signal for the LPC speech synthesizer at the receiver (see Fig. 5). In comparison with the speech amplitude spectrum (see Fig. 6), the output of the LPC analysis is spectrally flat. Encoding of a spectrally flat signal offers many avenues for efficient coding, as discussed in the following section.

LPC Synthesis

The LPC synthesis system is an inverse of the LPC analysis system. Thus, the synthesized speech in terms of the pitch-filtered prediction residual is

$$x(z) = \frac{1}{[1 - A(z)][1 - P(z)]} E_L(z), \tag{7}$$

where the quantity $\frac{1}{[1 - A(z)][1 - P(z)]}$ is the transfer function of the LPC synthesis filter. From Eqs. (6) and (7) it can be seen that the combination of the LPC analysis system and LPC synthesis filter

is a unity gain system even if *A(z)* and *P(z)* coefficients are quantized. Thus, the synthesized speech quality is solely dependent on the nature of quantization for the pitch-filtered prediction residual. We discuss this aspect in the next section.

Speech Parameter Quantization

The VDR voice processor generates slow time-varying parameters and fast time-varying parameters. The slow time-varying parameters include 10 prediction coefficients, a pitch prediction coefficient, an average or peak amplitude coefficient, a pitch period, and a loudness coefficient, and they are transmitted once every frame (22.5 ms or 180 speech sampling time intervals). The DOD narrowband voice terminals (ANDVT and STU-III) use the same speech data. In order to make our VDR voice encoder directly interoperable with ANDVT, we quantized these parameters in the manner as the DOD narrowband vocoders. Thus,

- The 10 reflection coefficients, representing speech resonant frequencies of each frame (22.5 ms), are quantized into 41 bits in the same manner as those of ANDVT or STU-III [3].
- The amplitude parameter, representing loudness of each frame, is quantized logarithmically into a 5-bit quantity in the same manner as that of ANDVT or STU-III [3].
- Pitch period is a consecutive integer number from 16 to 128. Thus, they are already quantized to a 7-bit quantity. Pitch periods of 15 and 128 correspond to pitch frequencies of 500 and 62.5 Hz.
- Pitch correlation coefficient is a long-term prediction coefficient at one period part. It is quantized into a 3-bit quantity, 8 levels from 0.3 to 1.0 spaced quasi-logarithmically. Exact breakpoint locations are not critical to output speech quality as long as they are spaced nearly equally.

Table 1 — Slow Time-Varying Speech Parameters

Slow Time-Varying Speech Components	Number of Bits Per Frame
10 Reflection Coefficients	41 bits
Amplitude Parameters	5
Pitch Period	7
Pitch Gain	3
TOTAL	56 bits (2.49 kb/s)

Excitation Signal Quantization

The pitch-filtered prediction residual is the output of the fast time-varying speech parameters generated by the LPC analysis system. Upon quantization, they become the excitation signal for the speech synthesizer at the VDR receiver. These excitation signal samples are quantized at a variable data rate from 2.93 to 32.27 kb/s. Hence, they are critical parameters for the VDR voice processor.

There are two generic ways of quantizing the excitation signal: open-loop quantization and closed-loop quantization. We have to choose one of them for our VDR voice processor.

Open-Loop vs Closed-Loop Quantization

Figures 9(a) and 9(b) show open-loop and closed-loop quantizers, respectively. In open-loop quantization, the quantizer is in series with other functional blocks. Thus, the quantized output is solely dependent on its input. On the other hand, in closed-loop quantization, the quantizer is within a closed loop. Therefore, its output is dependent on its input as well as its past output samples. We examine tradeoffs between these two methods of quantization in view of the VDR voice processor implementation.

The block diagram of a closed-loop quantizer can be derived from the block diagram of an open-loop quantizer by manipulating blocks. The transfer function of the open-loop system is

$$H(z) = [1 - A(z)][1 - P(z)], \tag{8}$$

which may be arranged as

$$[1 - A(z)][1 - P(z)] = \frac{1 - A(z)}{1 + \dfrac{P(z)}{1 - P(z)}}, \tag{9}$$

in which the LPC analysis filter, *1-A(z)*, may be rearranged as

$$1 - A(z) = \frac{1}{1 + \dfrac{A(z)}{1 - A(z)}}. \tag{10}$$

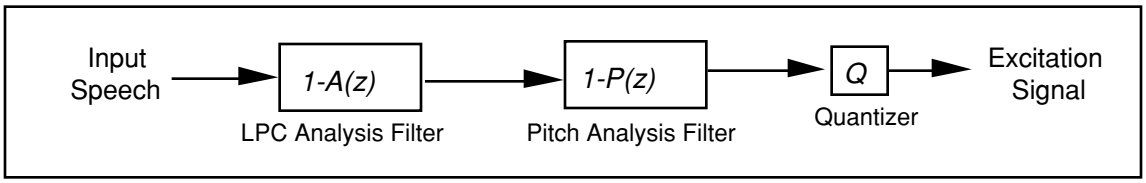
Substituting Eq. (10) into Eq. (9) gives

$$[1 - A(z)][1 - P(z)] = \left[1 + \frac{A(z)}{1 - A(z)}\right]\left[\frac{1}{1 + \dfrac{P(z)}{1 - P(z)}}\right], \tag{11}$$

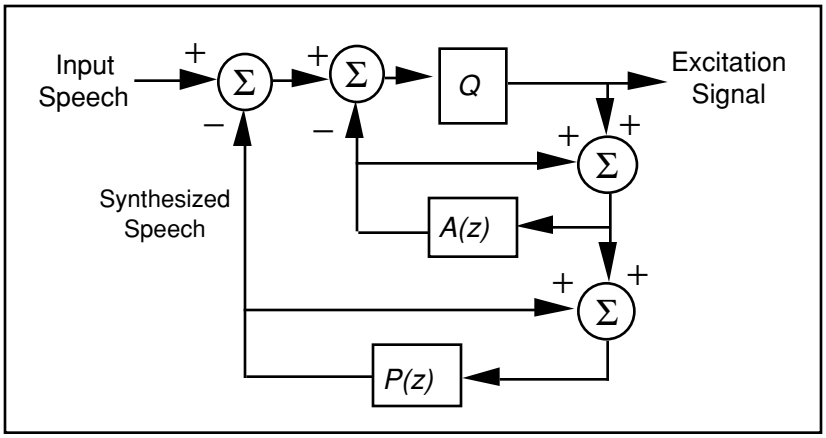
which can be represented by the block diagram shown in Fig. 9(b).

A closed-loop excitation signal quantizer has been known as the Adaptive Predictive Coder (APC) [4]. There are many variations of APC. One example is the Codebook-Excited Linear Predictor (CELP) operating at 4.8 kb/s used in STU-III. It was also used in an earlier DOD wideband vocoder APC operating at 16 kb/s and in another, Goldwine, operating at 6.4 kb/s.

A closed-loop excitation quantizer tries to minimize the difference between the input and output speech waveforms. Such a design principle is good for very high data rate, but it is not efficient for the medium data rates (9.6 to 20 kb/s) at which the VDR voice processor is operating. The output speech waveform can be very different from the input speech waveform, yet the output speech could sound very similar to the input waveform. Therefore, we dismissed the closed-loop quantizer for our VDR voice processor.



(a) Open-loop quantization of the excitation signal



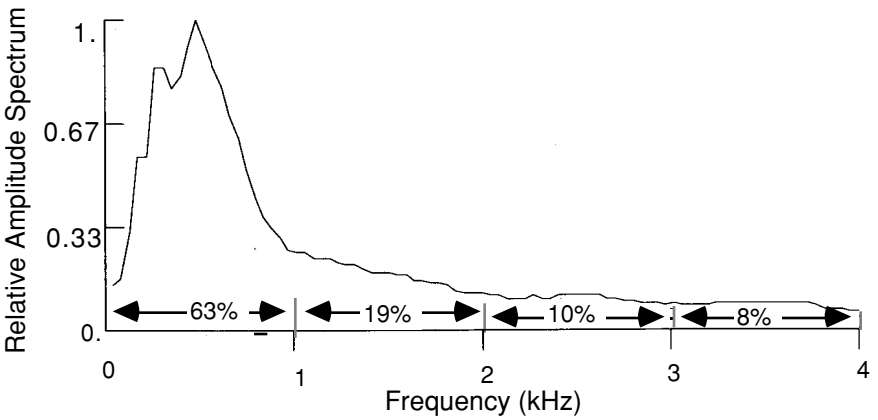
(b) Closed-loop quantization of the excitation signal

Fig. 9 — Open-loop and closed-loop quantization of the excitation signal. The open-loop quantization shown in Fig. 9(a) is a direct consequence of Eq. (4). The closed-loop quantization shown in Fig. 9(b) may be derived by manipulating blocks of Fig. 9(a), as indicated by Eqs. (8) through (11). The box indicated by Q is a quantizer, which rounds off the input signal amplitude into a finite number of steps.

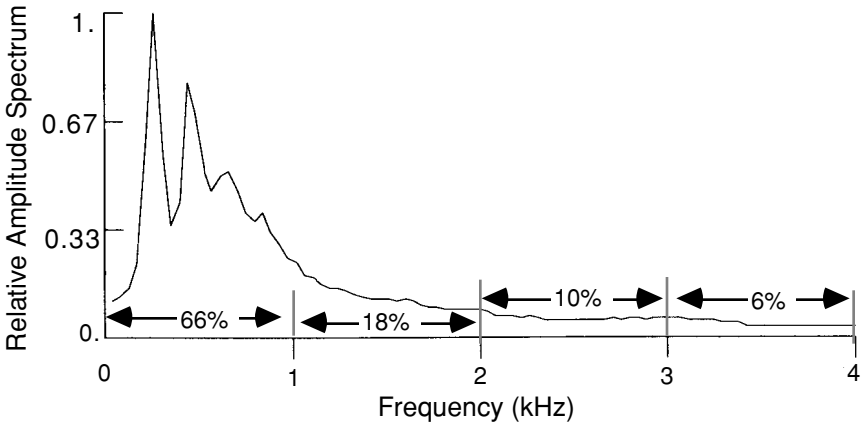
Time-Domain Quantization vs Frequency-Domain Quantization

Once the open-loop excitation quantization approach (see Fig. 9(a)) is selected, a choice must be made between time domain or frequency domain in terms of spectral components. In the time-domain quantization, each excitation bit represents the entire frequency contents equally. There is simply no way of adding more resolution for low-frequency contents that are perceptually much more significant than higher frequency contents. Therefore, we decided to use the frequency-domain approach, which has many advantages as discussed below:

1. *More Resolution for Low-Frequency Components*: According to our analysis of a variety of speech waveforms, approximately 65% of speech energy is concentrated in the 1-kHz band from 0 to 1 kHz (Fig. 10). In other words, speech contents from 0 to 1 kHz are perceptually more significant and, thus, need finer quantization. The frequency-domain encoding of the excitation signal makes it possible to exploit this important finding.



(a) 54 males are uttering two sentences each (7 min. 40 s)



(b) Twelve females are uttering two sentences each (2 min. 20 s)

Fig. 10 — Average speech spectrum with the percentage of speech energy within each 1-kHz bandwidth. Approximately 2/3 of the total speech energy lies between 0 and 1 kHz for either male or female speeches. Since a majority of speech energy is concentrated below 1 kHz, those spectral components should be quantized more accurately. Frequency-domain quantization makes the frequency-dependent equalization easy.

2. *Omission of Very Low Frequencies to Save Bits:* Spectral components below 130 Hz (four spectral components including DC component) need not be transmitted because we cannot hear them. We can save up to 1.6 kb/s (= $4 \times 9 \times 44.44$).

3. *Less Resolution for Higher Frequency Components:* It has long been well known that the human ear gradually loses frequency resolution capability for higher frequencies, as shown in Fig. 11. Thus, we allow coarser quantization for higher frequency spectral components.

4. *Spectral Replication to Save Bits:* For an average data rate of 20 kb/s, we are able to transmit all the spectral components. For lower average data rates, however, we don t have enough bits to transmit all spectral components. We have to do one of the following two steps:

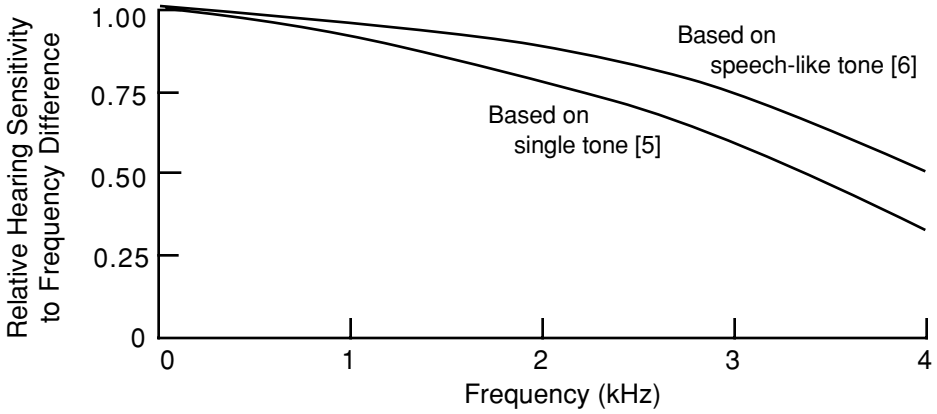


Fig. 11 — Relative hearing sensitivity to frequency differences. A speechlike tone has three resonant frequencies, which are repetitive. There are some differences in perception between single tone and speechlike tone, but the hearing sensitivity to frequency difference falls with frequency.

- Reduce spectral resolutions and transmit all spectral components.
- In lieu of reducing spectral resolutions, reduce the number of spectral components transmitted. At the receiver, the missing spectral components are regenerated by replicating the lower spectral components that are transmitted (see Fig. 12).

According to our experimentation, the second choice is far preferred over the first choice. The spectral replication method, originally conceived by this author, has been used to implement the Multirate (voice) Processor (MRP) deployed by the Navy [7].

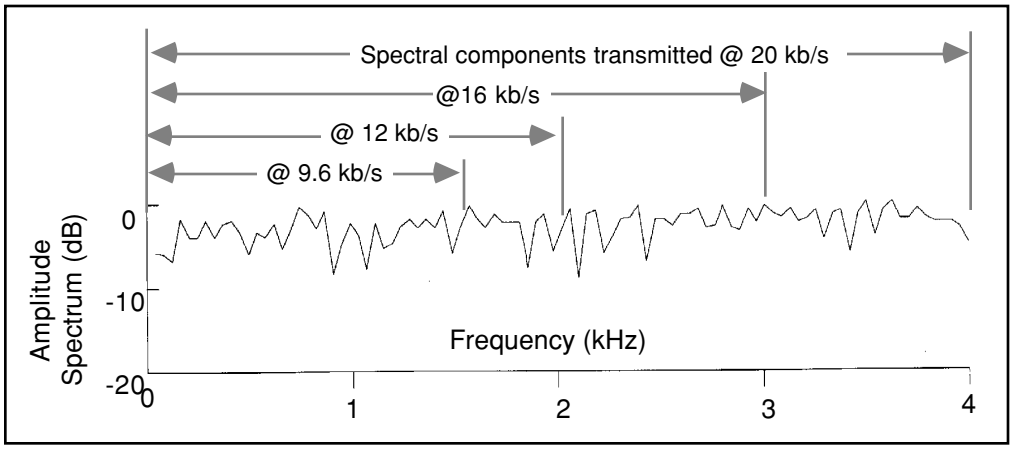


Fig. 12 — Spectral components transmitted for various average data rates. Some of the higher frequency components can be omitted from transmission to save bits. At the receiver, the missing higher frequency components are replaced by transmitted lower frequency components. Whenever an amplitude component is transmitted, the corresponding phase component (not shown here) must be transmitted jointly. This method works because the amplitude spectrum is relatively flat.

5. *Two-Dimensional Coding to Achieve Amplitude-Dependent Phase Resolution*: Because of spectral quantization of the excitation signal, the forward and inverse Fourier transforms tend to produce waveform discontinuities at the frame boundary. Thus, it is necessary to overlap frames at the expense of transmission efficiency (i.e., more bits are required to encode the same amount of excitation signal samples). We used an overlap size of 12 samples, making a Fourier transform size of 192 rather than 180. We used a 192-point real Fourier transform or 96-point complex Fourier transform [8]. We quantized both amplitude and phase parts jointly as a vector using a constellation of a unit circle that has a prescribed number of vectors. A 9-bit quantizer contains 512 points within a constellation, an 8-bit quantization contains 256 points, and so on. Figure 13 is an example of a 6-bit quantizer. One advantage of using two-dimensional coding of a frequency component is that we can allow coarser phase resolution of a low-amplitude component because that frequency component is less audible.

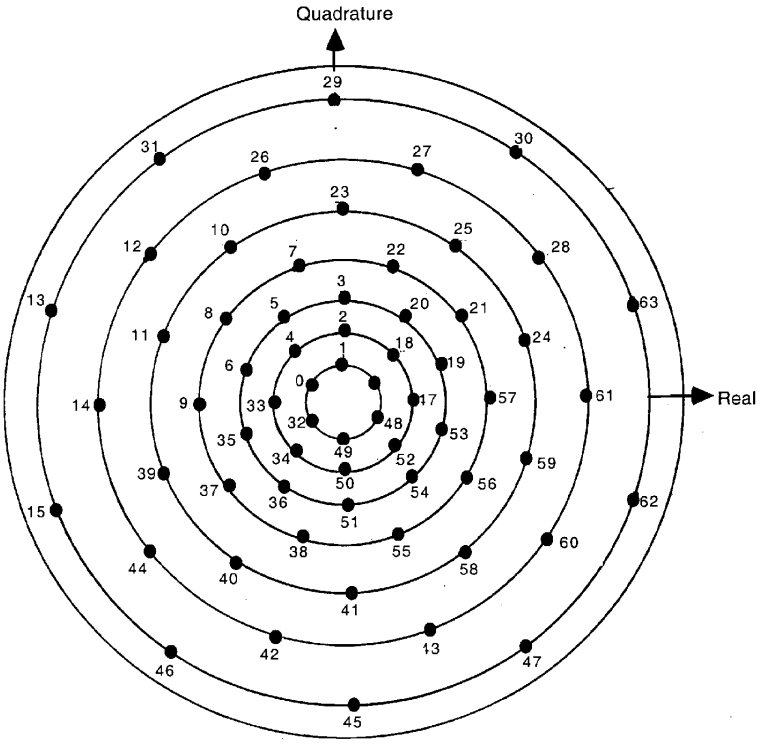


Fig. 13 — Constellation of 64 vectors to encode the excitation signal spectral components into a 6-bit quantity. We have a similar constellation for 3, 4, 5, 6, 7, 8, and 9-bit quantizers.

A Parameter that Indicates Speech Waveform Complexity

To implement a VDR voice processor, we need a parameter that indicates the complexity of the speech waveform so that we can select an appropriate quantization step to generate variable data rates. In the LPC analysis, a speech sample is represented by a weighted sum of the past 10 speech samples (see Eq. (1)). Thus, if the speech waveform is complex, the LPC analysis becomes less effective, thus producing larger errors. Therefore, the magnitude of prediction errors is a reliable indication of the speech waveform complexity. We partitioned the total dynamic range of the residual into seven logarithmic steps in such a way that the percentage of amplitude hits is nearly equal in all amplitude steps. Based on this amplitude step (see Table 2), we assigned the number of bits for each spectral components transmitted. In assigning bits, we incorporated the human perception characteristics shown in Fig. 11.

Table 2 — Index that Indicates the Speech Waveform Complexity and Spectral Resolution of the Excitation Signal

Spectral magnitude of Pitch-Filtered Prediction Residual	Waveform Complexity Index (A)	Spectral Resolution of Excitation Signal		
		0 - 1 kHz	1 - 2 kHz	2 - 4 kHz
359 - 511	Complex 1	9 bits	8 bits	7 bits
148 - 300	2	8	7	6
73 - 147	3	7	6	5
36 - 72	4	6	5	4
18 - 35	5	5	4	3
5 - 17	6	4	3	-*
0 - 4	Simple 7	3	-*	-*

* These spectral components are not transmitted because their contribution to synthesized speech is insignificant. At the receiver, random amplitude and phase spectral components are used for these.

Bit Assignments for Four Average Data Rates

Based on the spectral quantization rules listed in Table 2, the pitch-filtered prediction residual spectral samples are quantized as listed in Table 3. As mentioned previously, for lower average data rates, some of the higher spectral components are not transmitted. They are substituted at the receiver by low-frequency components. Regeneration of high-frequency excitation signals from low frequencies works well if the low-frequency components cover at least 0 to 1 kHz (in our case, 0 to 1.5 kHz). The resultant high-frequency signals are acceptably good. This form of high-frequency regeneration works in this application because the pitch-filtered prediction residual is spectrally flat.

Table 3 — Bit Assignments for the Excitation Signal for Four Average Data Rates

(a) Average data rate = 20 kb/s

A	0 to 1 kHz			1 to 2 kHz			2 to 3 kHz			3 to 4 kHz			Instantaneous Data Rate (kb/s)*
	B	N	T	B	N	T	B	N	T	B	N	T	
1	9	22	198	8	24	192	7	24	168	7	24	168	34.755
2	8	22	176	7	24	168	6	24	144	6	24	144	30.577
3	7	22	154	6	24	144	5	24	120	5	24	120	26.400
4	6	22	132	5	24	120	4	24	96	4	24	96	22.222
5	5	22	110	4	24	96	3	24	73	3	24	73	18.044
6	4	22	88	3	24	72	-	24	-	-	24	-	9.600
7	3	22	66	-	24	-	-	24	-	-	24	-	5.422

* Instantaneous data rates are referred to as the 7 data rates available from each frame of 22.5 ms.

Legend: *A* is the index representing the speech waveform complexity defined in Table 2.
B is the number of bits for each spectral vector (amplitude and phase components).
N is the number of spectral vectors within the frequency range specified.
T is the total number of bits for spectral vectors within the frequency range specified.

Table 3 — Bit Assignments for the Excitation Signal for Four Average Data Rates (continued)

(b) Average data rate = 16 kb/s

A	0 to 1 kHz	1 to 2 kHz	2 to 3 kHz	3 to 4 kHz	Instantaneous Data Rate (kb/s)
1	Same as above (Embedded)	Same as above (Embedded)	Same as above (Embedded)	Not Transmitted*	27.288
2					24.177
3					21.066
4					17.955
5					14.844
6					9.600
7					5.422

* The 0 to 1 kHz spectral components are replicated for these high frequencies at the receiver.

(c) Average data rate = 12 kb/s

A	0 to 1 kHz	1 to 2 kHz	2 to 4 kHz	Instantaneous Data Rate (kb/s)
1	Same as above (Embedded)	Same as above (Embedded)	Not Transmitted*	19.822
2				17.778
3				15.773
4				13.689
5				11.644
6				9.600
7				5.422

* The 0 to 2 kHz spectral components are replicated for these high freqencies at the receiver.

(d) Average data rate = 9.6 kb/s

A	0 to 1.5 kHz	1.5 to 4 kHz	Instantaneous Data Rate (kb/s)
1	Same as above up to 1.5 kHz (Embedded)	Not Transmitted*	15.555
2			14.044
3			12.533
4			11.022
5			9.511
6			8.000
7			5.422

* The 0 to 1.5 kHz spectral components are replicated more than once for these high frequencies at the receiver.

Legend: *A* is the index representing the speech waveform complexity defined in Table 2.
B is the number of bits for each spectral vector (amplitude and phase components).
N is the number of spectral vectors within the frequency range specified.
T is the total number of bits for spectral vectors within the frequency range specified.

CD/ONLINE AUDIO DEMONSTRATIONS

Included on the accompanying CD are three audio clips to illustrate various aspects of the VDR voice processor. The clips are also posted on our web site (http://www.nrl.navy.mil/).

Audio Demo I: VDR Speech with Instantaneous Data Rates Shown

The speech heard has been processed by our VDR voice encoder in which the average data rate is 20 kb/s. Figure 14 shows the histogram of actual variable data rates. The most remarkable fact is that the processed speech is completely free from warbles and flutters, although instantaneous data rates have been changed as often as 44 times per second. Note the low data rates between phrases. We don t need silence detection to save bits. It is done automatically for the VDR voice processor.

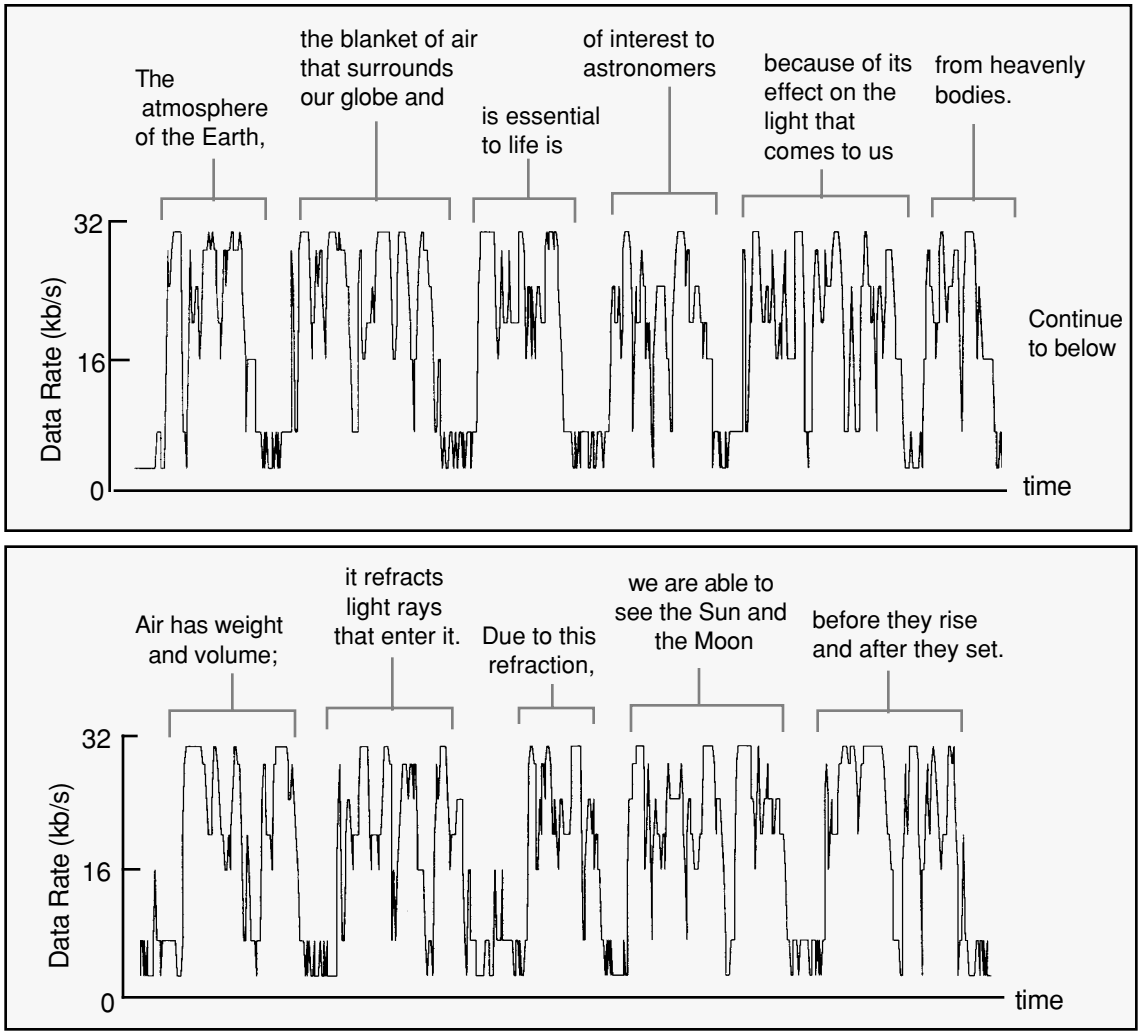


Fig. 14 — Data rate histogram obtained from the 20-kb/s mode of the VDR voice encoder. Voice data rate is actually variable as shown above. It is a remarkable fact that there are no audible transitions of data rates in the output speech. We have not seen an example of speech generated at variable data rates such as this. This speech sounds excellent.

Audio Demo II: Acoustic Noise Tolerance

The input speech sample is a tactical message with 20-mm machine gun noise in the background. As shown in the speech spectrum (Fig. 15), the machine gun noise is sporadic and extremely noisy (115 to 120 dB, sound pressure level). This is typical of noise at the front line. The spoken words are

Lima. This is Hotel. The enemy position 532714. The demo plays raw and processed speech samples in the following sequence:

1. Raw speech (for reference)
2. VDR at 20 kb/s
3. VDR at 16 kb/s
4. VDR at 12 kb/s
5. VDR at 9.6 kb/s
6. Speech encoded at 2.4 kb/s by a DOD vocoder (for reference)

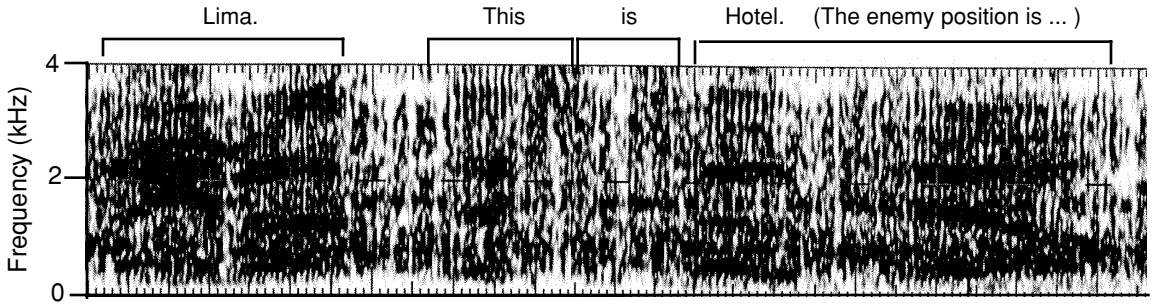


Fig. 15 — Spectrogram of part of test phrase. Background noise is 20-mm machine gun noise. Our VDR voice encoder provides good speech quality at any of four average data rates. Many voice encoders are not designed to operate in extremely noisy environments. As heard in this audio demo, a voice data rate of 2.4 kb/s does not work well in this kind of operating environment.

Audio Demo III: Switching of Data Rates on the Fly

The input speech sample is a local AM broadcast with fast talking for 37 seconds. While continuous speech is spoken, average voice data rate is switched once every second (see Fig. 16). This audio demo shows that the average data rate can be changed on the fly without introducing clicks, pops, or any other acoustic impediments. Speech quality degrades only slightly as the average data rate decreases from 20 kb/s to 9.6 kb/s. The data rate at the beginning of the speech sample is 20 kb/s.

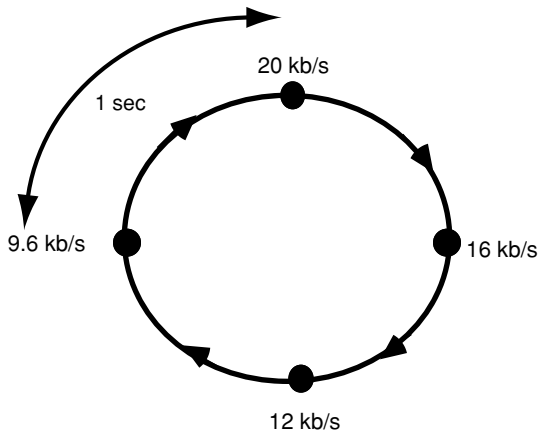


Fig. 16 — Average data rate change every second during continuous speech. No clicks or pops are heard during speech.

CONCLUSIONS

In the early 1980s, DOD tested various 16 kb/s voice processors in an attempt to develop a very high-quality secure phone for use in its office environments. We still remember somewhat raspy speech quality even at that high data rate. Now we know what the trouble was. Simply stated, a constant data rate of 16 kb/s is not sufficient to reproduce high-quality vowel sounds, although consonant sounds are good at 16 kb/s. We must use a higher data rate for vowels but we can use a much lower rate for consonants, and an even lower rate for gaps and silence. We need an *average* data rate of 16 kb/s, not a *fixed* data rate of 16 kb/s.

With the arrival of the VoIP in DOD, we will be able to transmit speech at variable data rates. This report presents, for the first time, a VDR voice processor that automatically alters data rate based on the complexity of the speech waveform. The quality of VDR speech is excellent at an average data rate of less than half to one-third of a fixed-data-rate speech. Behind that excellent quality speech, however, the data rate continuously changes (as often as 44 times per second); yet, there is absolutely no flutter or wobble in speech. This must be heard to be believed.

The characteristics of our VDR voice processor, listed in the Introduction, are needed for operation in military environments. We are in the process of incorporating our VDR voice processor into NRL s IP voice terminal with a gateway that makes it possible to interoperate with external legacy voice equipment. This IP terminal with our VDR voice processor will participate in a Fleet exercise in the near future.

ACKNOWLEDGMENTS

This project was partially sponsored by NRL and partially sponsored by SPAWAR. The author thanks Dr. Randy Shumaker, Superintendent of NRL s Information Technology Division, who has been sponsoring our R&D efforts for many years.

The author also thanks Vanessa Hallihan, Program Manager of the Navy Information Security Office (SPAWAR PMW-161), who promoted the project documented in this report. The author also expresses his appreciation to Mike Weber and Bill Kordela of the same organization who supported our R&D tasks and provided technology migration avenues for our R&D products.

Finally, the author thanks Brian Adamson of NRL Code 5523 who read the author s manuscript and provided appropriate comments.

REFERENCES

1. U. Black, *Voice over IP* (Prentice Hall, New Jersey, 2000) p.1.

2. J.P. Maker and R.B. Adamson, IVOX - The Interactive VOice eXchange Application, NRL Report 9805 (1996).

3. Federal Standard 1015, Analog to Digital Conversion of Voice by 2,400 Bits/Second Linear Predictive Coding, General Services Administration, GSA Specification Unit, 7th and D Street, Washington, DC 20407 (1984).

4. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs, New Jersey, 1978).

5. P. Lodefoged, *Elements of Acoustic Phonetics* (The University of Chicago Press, Chicago and London, 1974).

6. G.S. Kang and L.J. Fransen, Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSPs), NRL Report 8857 (1985).

7. G.S. Kang and L.J. Fransen, Second Report of the Multirate Processor (MRP) for Digital Voice Communication, NRL Report 8614 (1982).

8. H. Silverman, An Introduction to Programming the Winograd Fourier Transform Algorithms (WFTA), *IEEE Trans. Acoustics, Speech, Sig. Proc.* **ASSP-25**(2), 1977.